重新思考智商：对智力测验的基础、局限性和社会政治影响的批判性分析

Ernesto Mascarell Perez
Independent researcher
mascarell1960@hotmail.com

摘要

在第一个实用性智力测验创建一个多世纪后的今天，关于人类智力应如何描述、智商测验是否真正衡量智力、以及如果不是，它们究竟衡量什么，科学界仍然存在显著的分歧。本文对智力心理测量的认识论、方法论和伦理基础进行了批判性和系统性的回顾。研究考察了四个核心问题领域：（1）**"智力"构念和*g*因子的性**质与稳定性；（2）测量工具固有的文化和环境偏见；（3）遗传性研究设计中的方法论缺陷；以及（4）这些测验在教育、职业和临床情境中应用的社会和伦理后果。本文认为，智商测验声称能够测量一种普遍的、先天的、可等级化的认知能力，这一主张缺乏坚实的实证基础，在很大程度上构成了一种统计和文化的产物。文章最后为未来研究和专业实践提出了建议，倡导采用替代性的人类潜能评估模式。

关键词：智商，智力测验，***g*因子，文化偏**见，遗传性，心理测量学，生物决定论，效**度**

1. 引言

对人类智力的测量构成了科学心理学中最雄心勃勃也最具争议性的项目之一。自阿尔弗雷德·**比奈和西奥多·西蒙于1905年开**发出第一个实用性量表以来，智力测验已在全球范围内普及，用于教育情境中识别"资优"**和"智力障碍"个体，在工作**场所用于人员选拔，在临床环境中用于诊断各种状况。然而，这种普遍性与关于它们所声称测量的构念的性质及其应用有效性的根本性辩论的持续存在形成了鲜明对比。

肯·理查森（2002）**清晰地**阐述了这一核心悖论：**"在我**们接近第一个实用性智力测验百年之际，关于人类智力应如何描述、智商测验是否真正衡量它、以及如果

不是，它们真正衡量什么，科学界仍然几乎没有共识。"这种不确定性不仅仅是学术上的技术性问题；它对社会的流动性、教育公平和分配正义具有深远的影响。

本文的目标是双重的。一方面，它旨在对质疑智力心理测量基本假设的实证证据提供批判性的综合。另一方面，它寻求审视应用这些测验的实际后果，认为对它们的不加批判的使用往往在科学中立的外表下强化了结构性不平等。

2. 构念问题：什么是智力？

2.1 *g*因子的具体化

由查尔斯·斯皮尔曼在20世纪初提出的一般智力或*g*因子的概念，仍然是当代大多数智商测验的理论支柱。斯皮尔曼观察到，不同认知测验的分数往往呈正相关——即所谓的"正相关矩阵"——并推断出存在一个潜在所有测验的共同因子。

古生物学家斯蒂芬·杰伊·古尔德在其著作《人的误测》（1981）中对这一推理进行了根本性的批判。古尔德认为，*g*因子是他所称的"具体化谬误"的产物：即倾向于将抽象概念转化为真实实体的倾向。测验之间的相关性并不能证明一般心理能力的存在；它同样可以由影响任何测验任务表现的方法论、经验或社会认知因素来解释。

2.2 CHC理论的问题

卡特-霍恩-卡罗尔（CHC）模型代表了系统化认知能力结构的最复杂的尝试。这个三层级的层次模型——顶端是*g*因子，中间层是多种广泛能力（如晶体智力Gc和流体智力Gf），底层是特定能力——已被广泛采纳为智力测验开发和解释的框架。

然而，即使其支持者也承认存在实质性的局限性。约翰·D·沃瑟曼（2019）指出，CHC模型存在"解释力薄弱、无法做出可检验的预测、与单一测量工具（伍德科克-约翰逊测验）纠缠不清、其因子结构的不一致独立复制，以及最重要的是，其对特定学习障碍学生的识别和干预的证明价值未经证实。"沃瑟曼补充说，这些局限性是任何主要通过因子分析开发或测试的心理学理论所固有的。

2.3 理论替代方案及其实证状况

与单一模型相对，多维度的提议已经出现。霍华德·加德纳的多元智能理论（1983）假设存在至少八个相对独立的认知领域，在教育界广受欢迎。然而，其科学地位存在问题。2023年发表在《心理学前沿》上的一篇综述得出结论："许多研究者声称加德纳的多元智能理论是一个神经迷思，因为他们没有看到支持其基于大脑的独立智能提议的证据。"因子研究并未证明智能是独立的，而对基于多元智能教学效果的研究也未探索积极效应的替代原因，且未采用标准科学方法进行。

就其本身而言，斯滕伯格的成功智力理论（1999）提出了一个三元模型（分析性、创造性和实践性）。斯滕伯格本人既批评了*g*因子理论家——因其使用受限范围的参与者、材料或情境背景——也批评了新型多元智能的捍卫者——因其很少或没有对其理论进行实证检验——主张采取一种承认智力多面性的中间立场。

这种碎片化的图景表明，远未就智力的性质达成共识，而是存在着多种具有不同实证支持程度的概念框架，其中没有一个可以被认为是决定性的。

3. 测量中的文化和环境偏见

3.1 弗林效应与历史可塑性

对于将智力视为固定和先天特质的观念来说，最令人不安的发现之一是弗林效应：整个20世纪智商分数的持续增长。1981年，心理学家詹姆斯·弗林观察到，近一个世纪以来智商分数一直在稳步上升，两代人之间出现了惊人的18分差异。经过仔细分析，弗林得出结论，原因在于文化：社会在处理更大、更抽象的思想方面变得更加聪明，并使人们变得更加聪明。

这一发现直接削弱了将智力视为固定和先天实体的观点。正如大卫·申克所指出的，"这一观察，再加上智商分数在个体内部也不是固定的有力证据，清楚地消除了智力是一种先天和固定实体的观念。"分数随时间的最大增长往往与智商测验中文化导向内容的减少有关，这表明即使设计为"无文化影响"的测验也并非完全如此。

3.2 视觉空间推理测验中偏见的证据

即使是被呈现为"文化公平"的测验，如瑞文渐进矩阵或卡特尔文化公平测验，也无法摆脱文化偏见。2022年的一项综合性综述得出结论："在所有视觉空间推理测验中出现了几个根深蒂固的文化假设，以至于使将这些测验视为本质上文化公平的观点失效，并使从种族群体之间的平均分数差异中得出明确结论变得不可能。"

2025年发表的一项包含225项研究的系统综述确定了导致认知能力测验群体平均分数差异的十个因素，包括：生物和环境因素、测验焦虑、态度、经验和刻板印象威胁。结果表明，虽然生物和环境因素与群体差异有关，但若干与测验设计和实施相关的因素也同样有关，"最显著的是刻板印象威胁、经验和领域接触的差异，以及自信心。"

3.3 激进批判：智商测验作为社会阶层的替代指标

肯·理查森将批判向前推进了一步，认为"智商分数的全部人群变异可以用一个社会认知-情感因素纽带来描述，这些因素差异性地使个体为测验的认知、情感和表现要求做好准备——实际上，测验是社会阶层出身的衡量标准，而非复杂认知能力的衡量标准。"理查森展示了这些因素如何能够解释通常被认为验证了智商测验的相关性证据，包括与教育成就、职业表现和基本认知任务的相关性。

这种批判并不意味着否认个体在认知能力上的差异，而是质疑智商测验能否有效衡量这些差异，以及观察到的差异反映的是稳定的潜在能力，而非机会、资源和经验方面的累积不平等。

4. 智力的遗传：方法论回顾

4.1 遗传性辩论

支持智商测验有效性的最常被引用的论据之一是其所谓的高遗传性，通常估计在50%到80%之间。2026年发表在《科学报告》上的一项研究报告称，智商的遗传性约为75%，遗传因素解释了智商与社会经济地位之间的大部分关联（69-98%）。

然而，对这些估计必须谨慎解释。遗传性是一个随情境变化的人口参数，并不意味着个体差异是以固定方式由基因决定的。事实上，一些收养研究表明，当对范围限制进行校正时，共享环境可能解释高达50%**的智商**变异。

4.2 基因-环境交互作用

关于基因-环境交互作用的研究增加了额外的复杂性。塔克-**德**罗布及其同事（2015）**的一**项元分析发现，社会经济地位与智商遗传性之间的交互作用在不同国家之间并不一致。虽然一些美国研究报告称遗传性随着社会经济地位水平的提高而增加，但这种模式"**在美国以外的研究**（**英国、欧洲、澳大利**亚）中普遍不存在。"**作者推**测，这种差异可能反映了"**美国低社会**经济地位个体比其他国家更差的智力发展条件。"

对德克萨斯和科罗拉多收养项目数据的分析未发现实质性的遗传性×**社会**经济地位交互作用，效应量估计为负值或接近零。这表明社会经济地位可能无法完全捕捉到调节儿童智商的环境近端方面。

4.3 范围限制问题

收养研究中一个关键的方法论问题是家庭环境的范围限制。收养儿童的家庭往往不成比例地属于中上社会经济地位，这意味着所研究的环境范围并不代表总人口的全部变异性。对此限制的校正表明，"**共享**环境可能解释高达50%**的智商**变异。"这严重质疑了遗传因素解释大部分智商变异的说法。

5. 实践和伦理意义

5.1 工作场所的预测效度

智商测验对工作绩效的所谓预测效度一直是其在人员选拔中使用的关键论据。然而，最近的证据质疑了这一说法。贝里、利文斯、张和萨克特（2024）**的一**项更新元分析揭示，"**一般心理能力**测验的效标效度因不适当的范围限制校正而被大大高估。"**他**们的结果得出这样的结论："**排除一般心理能力**测验通常对效度影响很小或没有影响，但大大减少了不利影响。"

2023年发表在《职业与组织心理学杂志》上的一篇批判性综述审视了关于在研究生和专业职业选拔中使用认知能力测验的五个关键观察：（1）测验结果存在不利影响的证据；（2）**在高容量**选拔过程中将认知能力测验置于早期位置的倾向；（3）**最近挑**战认知能力测验元分析效度的证据；（4）历史主要效度研究的弱点；以及（5）对差异性效度证据的概念上有缺陷的审查。作者得出结论："**在**这些情境中应用认知能力测验的做法具有强烈的潜在可能性，会维持和加剧在获得更高职业和专业方面社会不平等。"

5.2 临床和教育应用

许多国家使用智商测验来确定特殊教育资格、获得卫生服务和智力障碍诊断。这些决定对个体具有终身影响。然而，鉴于这些工具中记录在案的文化和环境偏见，存在一种真正的风险，即它们将来自不利背景的儿童错误地标记为"**残疾**"，**而来自特**权背景的儿童被识别为"资优"**——不是由于先天潜能的差异，而是由于机会的差异。**

5.3 替代方案：动态评估

针对静态测验（如传统智商测验）的局限性，出现了侧重于学习潜力而非过去成就的替代方案。与鲁文·**福伊**尔斯坦和罗伯特·**斯滕伯格的工作相关的**动态评估，强调"**学**习潜力而非过去的学习成就"。这种方法遵循测试-训练-**再**测试的格式，衡量个体从指导中受益并在帮助下改进的能力——这是一种比当前水平的静态测量更接近许多人所理解的"**潜力"的方法**。

6. 结论与建议

本文回顾的证据表明，智商测验远非测量普遍和固定认知能力的中性工具，而是文化和历史情境中的产物，其分数在很大程度上反映了教育机会、社会文化经验和物质资源方面的不平等。*g*因子与其**说是关于人类心智架构的发现，不如说似乎是一种统计产物，其表面上的稳定性掩盖了深刻的语境依赖性。

这并不意味着否认个体在认知能力上存在差异。而是意味着质疑智商测验能有效和公平地衡量这些差异的主张，并拒绝接受观察到的差异反映了可等级化的先天一般能力的推论。

基于这一回顾，提出以下建议：

1. 对研究者：放弃寻找单一的普遍*g***因子**， 转而采用多因素和情境化的模型。优先研究认知可塑性及其促进条件，而非测量所谓的固定特质。
2. 对专业人员（教育工作者、临床心理学家、选拔人员）：在解释智商分数时保持最大限度的谨慎，明确考虑被评估者的社会文化背景和工具的局限性。认真考虑用测量学习潜力的动态评估方法替代静态智商测验。
3. 对政策制定者：在没有严格证明其对特定人群的有效性及其优于偏见较少的替代方案的情况下，避免将智商测验作为高风险决策（特殊教育准入、工作选拔）的基础。
4. 对专业培训：在心理学、教育和人力资源管理的课程中纳入关于智力测验的历史、基础和局限性的批判性培训，促进一种既承认其潜在用途也承认其严重局限性的细致理解。

7. 透明度声明

作者声明与本文内容没有相关利益冲突。支撑本工作的研究完全基于已发表和可公开获取的来源，且未获得特定资助。

参考文献

1. Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101(2), 171-191.
2. Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. Basic Books.
3. Gould, S. J. (1981). The mismeasure of man. W. W. Norton & Company.
4. Herrmann, E., et al. (2012). Are there geniuses among the apes? Philosophical Transactions of the Royal Society B, 367(1603), 2753-2761.
5. Hunt, E., & Carlson, J. (2007). Considerations relating to the study of group differences in intelligence. Perspectives on Psychological Science, 2(2), 194-213.
6. Jensen, A. R. (1980). Bias in mental testing. Free Press.
7. McGill, R. J., & Dombrowski, S. C. (2019). Critically reflecting on the origins, evolution, and impact of the Cattell-Horn-Carroll (CHC) model. Applied Measurement in Education, 32(3), 216-231.
8. Richardson, K. (2002). What IQ tests test. Theory & Psychology, 12(3), 283-314.

9. Richardson, K., & Norgate, S. (2015). Does IQ really predict job performance? Psychological Science, 26(2), 145-155.
10. Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. Holt, Rinehart & Winston.
11. Sternberg, R. J. (1999). Successful intelligence: Finding a balance. Trends in Cognitive Sciences, 3(11), 436-442.
12. Tucker-Drob, E. M., & Bates, T. C. (2016). Large cross-national differences in gene × socioeconomic status interaction on intelligence. Psychological Science, 27(2), 138-149.
13. Wasserman, J. D. (2019). Deconstructing CHC. Applied Measurement in Education, 32(3), 232-248.
14. Wilson, C. J., et al. (2023). The cross-cultural generalizability of cognitive ability measures: A systematic literature review. Intelligence, 98, 101751.
15. Woods, S. A., & Patterson, F. (2023). A critical review of the use of cognitive ability testing for selection into graduate and higher professional occupations. Journal of Occupational and Organizational Psychology.
16. Cuppello, S., Zibarras, L. D., & Corr, P. J. (2025). Factors related to mean score group differences on cognitive ability tests: A systematic review. Trends in Psychology.